Doble Grado en II+ADE 24/04/ 2018

Ecometría Examen Parcial Curso 2017-2018

DNI/NIE : APELLIDOS: Nombre:

Cada pregunta se resuelve en la hoja de su enunciado, no se pueden responder preguntas distintas en la misma hoja. Las respuestas se deben escribir con tinta azul o negra. Las respuestas deben ser breves pero razonadas. Errores conceptuales importantes pueden afectar a la calificación global del examen.

Teoría(3 puntos). 1) Enunciar el teorema de Gauss-Markov, tanto en regresión múltiple, como en regresión simple.

2) Demostrar el teorema de Gauss-Markov.

3) ¿Podría explicar la relevancia del teorema de Gauss-Markov?

1) y 2) están en los apuntes del curso. Observación: en 2) al no decir nada, se entiende que se debería demostrar tanto para el caso de regresión simple como múltiple, a partir del resultado para regresión múltiple (que es más completo), se obtiene el de regresión simple (Ejercicio II.10 de los apuntes).

3) De las tres propiedades más deseables en los estimadores, insesgadez, eficiencia (menor varianza o "menor"matriz de covarianzas en el caso de va multivariadas) y consistencia, el teorema hace referencia a las dos primeras respecto al estimador de mínimos cuadrados de los coeficientes. La consistencia es relevante para que los errores en las muestras sean los menores posibles. Debido a esto, se utiliza el acrónimo BLUE ("Best Linear Umbiased Stimator") para el estimador de mínimos cuadrados.

Doble Grado en II+ADE **24/04/ 2018**

Ecometría Examen Parcial Curso 2017-2018

DNI/NIE : **APELLIDOS:** **Nombre:**

Ejercicio(4 puntos). Dados los siguientes datos

y	x_2	x_3
4	5	2
6	4	3
4	5	2
5	7	2
4	4	2

1) Obtener el plano de regresión de y sobre x_2 y x_3. Dar un valor estimado de la varianza del estimador del coeficiente de la variable x_3.

2) ¿Son las variables explicativas conjuntamente significativas?

3) ¿Es cada una de las variables explicativas independientemente significativa?

4) Según lo anterior, ¿cúal es el modelo que representa más significativamente los datos?

Ayuda: $\begin{pmatrix} 5 & 25 & 11 \\ 25 & 131 & 54 \\ 11 & 54 & 25 \end{pmatrix}^{-1} = \begin{pmatrix} 359/19 & -31/19 & -91/19 \\ -31/19 & 4/19 & 5/19 \\ -91/19 & 5/19 & 30/19 \end{pmatrix} = \begin{pmatrix} 18.89 & -1.63 & -4.79 \\ -1.63 & 0.21 & 0.26 \\ -4.79 & 0.26 & 1.58 \end{pmatrix}$

1) Con los datos, construimos las matrices

$$X = \begin{pmatrix} 1 & 5 & 2 \\ 1 & 4 & 3 \\ 1 & 5 & 2 \\ 1 & 7 & 2 \\ 1 & 4 & 2 \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} 4 \\ 6 \\ 4 \\ 5 \\ 4 \end{pmatrix}.$$

Observamos que el método de mínimos cuadrados funcionará, ya que el rango de X es 3: tomar el menor formado con las filas 1, 2 y 4.

Entonces

$X'X = \begin{pmatrix} 5 & 25 & 11 \\ 25 & 131 & 54 \\ 11 & 54 & 25 \end{pmatrix}$. Utilizando la expresión de su inversa *mediante números racionales*, obtenemos

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1}X'\mathbf{y} = \begin{pmatrix} -40/19 \\ 7/19 \\ 42/19 \end{pmatrix}, \ y = -\frac{40}{19} + \frac{7}{19}x_2 + \frac{42}{19}x_3 \ \text{(plano de regresión)},$$

$$\begin{pmatrix}\hat{\varepsilon}_1\\ \hat{\varepsilon}_2\\ \hat{\varepsilon}_3\\ \hat{\varepsilon}_4\\ \hat{\varepsilon}_5\end{pmatrix} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = \begin{pmatrix}-3/19\\ 0\\ -3/19\\ 2/19\\ 4/19\end{pmatrix},\ s^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{2} = \frac{1}{19},\ s_2^2 = s^2\frac{4}{19} = \frac{4}{361},\ s_3^2 = s^2\frac{30}{19} = \frac{30}{361}$$

(aunque no piden s_2, lo hemos calculado, pues lo necesitaremos en el apartado 2). *La estimación de la varianza de* $\hat{\beta}_3$ *es* $s_3^2 = 30/361 = 0.0831$.

2) El test de hipótesis que planteamos es

H0: $\beta_2 = \beta_3 = 0$, H1: alguno de los parámetros β_2 o β_3 es no nulo.

Entonces, hemos de calcular el F-valor asociado a este test global, comparando el modelo con la hipótesis H0 (modelo constante) con el modelo completo con las dos variables explicativas:

$$SST = \sum(y_i - \bar{y})^2 = \frac{16}{5},\ R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{SST} = \frac{147}{152} = 0.9671,\ F = \frac{n-k}{k-1}\frac{R^2}{1-R^2} = \frac{R^2}{1-R^2} = \frac{147}{5} = 29.4.$$

Mirando la tabla de $F(k-1, n-k) = F(2,2)$, como el valor crítico para una signicación del 5 % es de 19, y $29.4 > 19$, rechazamos la hipótesis H0 y *las variables* x_2 *y* x_3 *son conjuntamente significativas.*

3) Aquí tenemos dos tests de significación independientes, uno para cada variable, comparando el modelo completo con el restringido al quitar una de las variables:

a) Test variable x_2: H0: $\beta_2 = 0$, H1: $\beta_2 \neq 0$.

b) Test variable x_3: H0: $\beta_3 = 0$, H1: $\beta_3 \neq 0$.

Aunque se podría realizar mediante un test F, es más simple con un test t, calculando los t-valores:

$$t_2 = \frac{\hat{\beta}_2}{s_2} = \frac{7}{5} = 3.5,\ t_3 = \frac{\hat{\beta}_3}{s_3} = \frac{42}{\sqrt{30}} = 7.668.$$

Mirando la tabla de $t(2)$ con dos colas, la región de aceptación es el intervalo $(-4.303, 4.303)$; como t_2 está en esa región y t_3 no lo está, no podemos rechazar H0 en el test a), rechazando H0 en el test b), es decir, *la variable* x_3 *es significativa, pero no lo es la variable* x_2.

4) Quitando variables, tenemos cuatro modelos posibles: el constante, los dos restringidos quitando una de las variables, y el completo. Por el apartado 2, el constante no es significativo respecto al completo y por el apartado 3) podemos prescindir de la variable x_2 en el completo: *el modelo que representa màs significativamente los datos es el modelo restringido con la única variable explicativa* x_3.

Observación: si en lugar de utilizar en $(X'X)^{-1}$ su expresión mediante números racionales, se utiliza su aproximación mediante decimales, entonces los errores de redondeo se acumulan dando resultados que en algunos casos, como en el del coeficiente de determinación, son manifiestamente erróneos (!). Como dábamos como expresión posible a utilizar la que contenía decimales, se han evaluado como correctos los resultados basados en ella, si el resto del procedimiento es el adecuado.

Doble Grado en II+ADE **24/04/ 2018**

Ecometría Examen Parcial Curso 2017-2018

DNI/NIE : **APELLIDOS:** **Nombre:**

Ejercicio con ordenador A. (1.5 puntos)(simulación de un modelo de regresión) Sea el modelo

$$y_i = 1 - 2x_{2i} + 3x_{3i} + N(0,2),\ i = 1, ..., 500$$

siendo

$\mathbf{x}_1' = (1, 1..., 1, 2, ..., 2, ..., 10, 10, ..., 10)$, repitiéndose 50 veces cada uno de los dígitos del 1al10,

$\mathbf{x}_2' = (1, 1..., 1, 2, ..., 2, ..., 25, 25, ..., 25)$, repitiéndose 20 veces cada uno de los dígitos del 1al25,

1) Admitiendo la independencia de las perturbaciones, ¿cumple el modelo anterior las hipótesis de un modelo de Gauss-Markov? (Observación: aunque realmente no es necesario, el programa R calcula determinantes de matrices, mediante el comando "det")

2) Generando una tabla de datos del modelo, proceder al ajuste del mismo. Comentar los resultados: estimación de los parámetros, plano de regresión, significación de los modelos.

3) ¿Eran de esperar los resultados anteriores?.

1) Por construcción, cumple las hipótesis

(H1) (linealidad:por la fórmula que nos dan),

(H3) la esperanza de la perturbación es cero al ser $\varepsilon_i = N(0,2)$,

(H4) la varianza de la perturbación es constante $= 4$,

(H5) y (H6) al decirnos que las perturbaciones son independientes. Falta la (H2) (colinealidad), es decir, que el rango de la matriz X sea 3. Pero tomando un menor adecuado de la matriz X (aquí podría ser conveniente visualizar los datos), por ejemplo,

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{vmatrix} \neq 0 \Rightarrow \text{rang}(X) = 3.$$

Por tanto, se cumplen todas las hipótesis (H1)-(H6).

2) Generando una tabla de datos y ajustando el modelo mediante regresión lineal, redondeando a dos decimales, obtenemos:

Estimación de los coeficientes: $\hat{\beta}_1 = 1.17$, $\hat{\beta}_2 = -2.19$, $\hat{\beta}_3 = 3.07$ y el plano de regresión es $y = 1.17 - 2.19x_2 + 3.07x_3$.

Estimación de la desviación típica, σ: $s = 1.88$.

Observamos que las estimaciones de los parámetros son cercanas a sus valores reales.

La significación global es muy buena: p-valor del test F de significación global nulo a efectos prácticos, con lo que rechazamos con rotundidad H0: $\beta_2 = \beta_3 = 0$. La significación de los dos

test t individuales de cada una de las variables explicativas x_2 y x_3 también es muy buena, con p-valores también muy pequeños (rechazamos las hipótesis H0: $\beta_2 = 0$ y H0: $\beta_3 = 0$). Sin lugar a dudas, el modelo completo es el adecuado.

3) Los resultados anteriores eran de esperar, pues hemos partido de un modelo estadístico de Gauss-Markov (el llamado "modelo real", que usualmente es desconocido en las aplicaciones prácticas) para generar los datos.

Observaciones. Para generar los datos se han utilizado en la consola de R las instrucciones

$> x2 <-$ as.numeric(gl(10,50))

$> x3 <-$ as.numeric(gl(25,20))

$> y <- -1 - 2 * x2 + 3 * x3$+rnorm(500,mean=0,sd=2)

$>$ datosA $<-$ data.frame(y,x2,x3)

entonces en la consola de R-Commander, se busca el fichero de datos (pestaña "conjunto de datos"), en nuestro caso, "datosA".

Generando diversas tablas de datos, los resultados no son necesariamente idénticos, pero serán similares. En particular, la significación de los tests será muy buena en todos ellos.

Doble Grado en II+ADE **24/04/ 2018**

Ecometría **Examen Parcial** **Curso 2017-2018**

DNI/NIE : **APELLIDOS:** **Nombre:**

Ejercicio con ordenador B. (1.5 puntos) Basándonos en el fichero de datos longley.csv (ver también el fichero longley.docx), vamos a estudiar el comportamiento del empleo (Employed) en una población estadounidense a lo largo de los años 1947 to 1962, repecto a tres variables explicativas: deflactor del PNB ("GNP.deflator", tomando 1954 como año de referencia), población ≥ 14 años ("Population") y paro ("Unemployed"). Este es el modelo que consideraremos como completo. (Observación: al importar el fichero, hay que marcar que la separación de campos es mediante comas.)

1) Escribir los resultados de la regresión del modelo completo: plano de regresión, estimación de los parámetros, coeficiente de determinación, errores estándar, t-valores, p-valores, etc. Observamos que uno de los coeficientes del plano de regresión es negativo, ¿cómo se interpreta esto?

2) A la vista de los resultados anteriores, ¿son conjuntamente significativas las variables explicativas?, ¿lo son individualmente?. ¿Cúal es el modelo significativamente más relevante entre los obtenidos a partir de los resultados de 1)?. Interpretación.(Observación: no se pide estudiar la significación de todos los modelos restringidos posibles quitando variables explicativas, sino sólo de los que se pueden analizar mediante los resultados del apartado 1).

3) A partir de los resultados de 1), se observa que el cociente de dos de los coeficientes del plano de regresión es aproximadamente -38. Entonces es razonable plantear la hipótesis de que el cociente de los correspondientes parámetros del modelo estadístico tiene ese valor, ¿es esta hipótesis significativamente relevante?

1) Por comodidad, denotaremos las variables explicativas deflator del PNB, población y paro por x_G, x_P y x_U, respectivamente.

Ajustando el modelo mediante regresión lineal, redondeando hasta 5 decimales debido a que el paro tiene datos con decimales significativos a partir de 10^{-3}, obtenemos

	$\hat{\beta}_j$	s_j	t_j	p-valor
intercepto	11.52468	7.37085	1.564	0.14390
x_G	0.13669	0.07763	1.761	0.10373
x_P	0.36587	0.12988	2.817	0.01555
x_U	-0.00961	0.0025	-3.838	0.00236

$s = 0.6174$, $R^2 = 0.9753$, $\bar{R}^2 = 0.9691$,

test F de significación global con $F(3, 12) : F - \text{valor} = 157.8$, $p - \text{valor} = 6.621 \times 10^{-10}$.

El plano de regresión es $y = 11.52468 + 0.13669x_G + 0.36587x_P - 0.00961x_U$. El coeficiente del paro es negativo, lo cual es razonable céteris páribus: manteniéndose todas las otras variables constantes. a mayor desempleo, menor empleo.

2) Con los resultados anteriores se pueden analizar cuatro modelos comparando con el completo:

a) los tres restringidos al prescindir de una de las variables explicativas, mediante 3 tests t:

H0: $\beta_j = 0$, H1: $\beta_j \neq 0$, j = G, P, U

b) el modelo constante, mediante el test de significación global:

H0: $\beta_G = \beta_P = \beta_U = 0$, H1: algún $\beta_j \neq 0$.

Basándonos en 1), al 5 %, como el p valor asociado al test b) es menor que 0.05, rechazamos H0 en el test b): el modelo constante no es significativo, respecto al completo. Análogamente, como los p valores del test t son menores que 0.05 para las variables x_P y x_U rechazamos H0 en los test a), pero no la rechazamos para la variable x_G: esta variable no es significativa cara al empleo. Por tanto, el modelo significativo para explicar el empleo es el de las variables x_P y x_U. En particular, la inflacción no tiene mucha influencia sobre el empleo.

3) Efectivamente, mirando los resultados de 1),

$$\frac{\hat{\beta_P}}{\hat{\beta_U}} = -38.072 \approx -38.$$

Por tanto, lo que queremos estudiar es el contraste con un test F:

H0: $\beta_P + 38\beta_U = 0$, H1: $\beta_P + 38\beta_U \neq 0$.

Para ello, construimos un nuevo modelo a partir del completo (ya que las estimaciones son del completo), definiendo una nueva variable z:

$$\beta_P = -38\beta_U \Rightarrow \beta_P x_P + \beta_U x_U = -38\beta_U x_P + \beta_U x_U = \beta_U(x_U - 38x_P) = \beta_U z,$$

siendo $z := x_U - 38x_P$.

Definiendo en R-Commander esa nueva variable (Datos->Modificar variables del conjunto de datos activo->Calcular una nueva variable), se obtiene un nuevo conjunto de datos con esa variable adicional. Se ajusta la regresión de esos datos con las variables explicativas x_G y z y se comparan ambos modelos de regresión, el completo y este último (Modelos->Test de hipótesis->Comparar dos modelos). Del resultado, sólo nos interesa el p-valor, que es 0.9954. Por tanto, claramente no se puede rechazar H0 y consecuentemente *la hipótesis de que* $\hat{\beta_P}/\hat{\beta_U} = -38$ *es significativa.* No lo piden, pero fijémonos que esto significa que, céteris páribus, el efecto de la población sobre el empleo es del orden de -38 veces el del paro; el signo negativo significa que una variable tiende a aumentar el empleo y la otra a reducirlo.